LLNL Environmental Restoration Division (ERD) Standard Operating Procedure (SOP)	
ERD SOP 5.21: Outlier Identification Program	
REVISION: 0	AUTHOR: D. MacQueen REVIEWERS: V. Dibley, G. Kumamoto, T. Ottesen, and M. Ridley
EFFECTIVE DATE: June 1998	Page 1 of 21
APPROVAL Date Patricia Ottesen 5-20-98 Information Systems Management Group Leader	APPROVAL Date Albert J Lamarre 6-11-98 Division Leader CONCURRENCE Date Valerie D Dibley 5/20/98 QA Implementation Coordinator

1.0 PURPOSE

This procedure describes the Outlier Identification Program and defines the statistical outlier identification process. The process includes: identifying potential outliers, reviewing the outlier program output, and recording outlier status in the Environmental Protection Department database EPDData. Subsequent retrievals from EPDData may then include outlier status, in order to warn data users of anomalous results.

2.0 APPLICABILITY

This procedure governs the use of the computer programs developed for the Outlier Identification Program.

The statistical method employed by this program is not the only possible way to identify outliers. Therefore, the existence of this procedure does not prevent ERD staff from using other methods for outlier identification that may be appropriate in other situations.

3.0 REFERENCES

Barnett, V. & T. Lewis (1994), *Outliers In Statistical Data*, 3rd ed. (Wiley & Sons).

4.0 DEFINITIONS

See SOP Glossary.

5.0 RESPONSIBILITIES

5.1 Outlier Custodian

The outlier algorithm custodian runs the SAS computer codes used to identify and review outliers. The outlier custodian should be familiar with the UNIX operating system, be able to use a UNIX text editor, and be familiar with the EPDData database and Structured Query Language (SQL) used to retrieve data from the database.

5.2 QC Chemist

The QC Chemist is responsible for reviewing the outliers and deciding whether or not to accept the outlier algorithm recommendation.

5.3 Statistician

The statistician is responsible for choosing the statistical methods used, assessing their performance, finding ways of improving the statistical performance, and explaining the program to users.

5.4 Database Programmer

The database programmer develops and maintains the RDBMS computer codes used by the algorithm.

5.5 Outlier Programmer

The outlier programmer develops and maintains the SAS computer codes used to identify and review outliers. The outlier custodian should have substantial skill and experience with the SAS software system, the UNIX operating system, the EPDData database, and SQL.

5.6 Project Staff

Project staff are the end users of the data, i.e., the individuals responsible for using environmental data to make programmatic decisions.

6.0 PROCEDURE

An outlier run consists of the following steps:

- Identify outliers
- Review outliers
- Upload, update, and append outliers

Outlier runs should be performed regularly, preferably at least quarterly. Each outlier run examines a set of analytes at a set of sampling locations. For example, an outlier run can examine all analytes in ground water (for which data is available) at the set of wells associated with Treatment Facility A.

6.1 File Locations

Data files, computer programs, and listings of results are maintained in the EPD UNIX network. The logical name of the outlier root directory is:

`/erd/statistic/project/outliers`.

This directory is available through automounting in the Wonderland domain. Runs for the Livermore Site are performed in subdirectory s200. Runs for Site 300 are performed in subdirectory s300.

6.2 Log Files

There is a log file for each site. For Site 300 the file is S3-outlier-log, in the s300 subdirectory. For the Livermore Site the file is `LS-outlier-log`, in the `s200` subdirectory.

It is absolutely essential that the Outlier Custodian maintain in these files a record of the steps performed.

6.3 Outlier Identification

Outliers are identified by a computer program written using SAS Software. This step is performed by the Outlier Custodian and are described in Sections 6.4 through 6.8.

- Retrieve data from EPDData.
- Perform a statistical analysis of each analyte at each location to identify outliers.
- Save a dataset containing new outliers and previously identified outliers that have changed status.
- Prepare data for review.

6.4 Prepare a Working Directory

6.4.1 Log in to a Sun SparcStation computer in the Wonderland domain. The computer must be one on which both SAS and OpenIngres are available. Currently, EPGEM is the best choice.

6.4.2 Move to the outlier program root directory, /erd/statistic/project/outliers. All subsequent references to directories will be relative to this directory. From there, move to directory s200 for a Livermore Site run, or s300 for a Site 300 run.

6.4.3 Check the log file to find out what area the previous run covered. Make sure that the previous run was completed, including the upload and append step (Section 6.13).

6.4.4 Decide what set of locations is to be evaluated. The set of locations will be defined by one or more values of the area2 field in the EPDData location table. For example, area2='EWFA', or area2 in ('TFG', '5475').

6.4.5 Choose an abbreviation, at most four characters long, that is suggestive of the set of locations chosen for this run. The abbreviation will be used to automatically generate file names during the outlier run. Examples of abbreviations that have been used are:

- ewfa East Firing Area/West Firing Area
- gsa General Services Area
- he High Explosives Processing Area
- pit6 Pit 6 Area
- b83 833 and 834 areas
- tfa Treatment Facility A (TFA)
- tfb Treatment Facility B (TFB)
- tfcd Treatment Facilities C and D (TFC & TFD)
- tfef Treatment Facilities E and F (TFE & TFF)
- tfg5 Treatment Facilities G and 5475 (TFG and 5475)

Note: These abbreviations may be reused.

6.4.6 Create a new directory. The directory name should include the abbreviation, and should uniquely identify this run relative to previous or subsequent runs for the same set of locations. For example, tfg5.97b is an appropriate name for the second outlier run in 1997 for locations associated with TFG and 5475. Before creating the directory, use the UNIX ls command to find out what subdirectories are present, and make sure the new name is unique. Follow the directory naming scheme indicated by the directories that are present.

Note: The new directory will be referred to as the working directory in the following sections.

6.4.7 Change to the working directory, and create a new subdirectory there named ssd.

6.5 Prepare Command Files

If necessary, change to the working directory.

6.5.1 In the remainder of this procedure, the string 'xxx' will be used to represent the abbreviation chosen in Section 6.4.5. Everywhere that 'xxx' appears, substitute the abbreviation.

6.5.2 Four SAS command files must be created in the working directory. One is always named autoexec.sas. The other three are named xxxrun.sas, xxxacc.sas, and xxxupl.sas.

6.5.3 The best way to create these four files is to copy them from the previous run. For example, if the previous run was in the directory ewfa.97a and the new working directory is gsa.97b, then copy ewfarun.sas, ewfaacc.sas, and ewfaupl.sas from directory ewfa.97a to directory gsa.97b, and rename them as gsarun.sas, gsaacc.sas, and gsaupl.sas respectively. Copy autoexec.sas also, but do not rename it.

6.5.4 Edit the four new files. In all four files, change the comments at the beginning of the file to reflect the current date and the set of locations in the outlier run. Comments are delimited by '/*' before the comment, and '*/' after the comment.

6.5.5 In addition, edit the files autoexec.sas and xxxacc.sas as follows:

1. autoexec.sas

6.5.6 In this file there are four SAS program statements that begin with '%let'. These statements define values of SAS macro variables that control certain aspects of each outlier run. Normally, only the variables ocode and title will need to be changed. Edit them as needed, as indicated here:

SAS macro variables.

Site 300	Livermore Site
%let site = s300;	**%let site = s200;**
%let wtable = uo3;	**%let wtable = uol;**
%let ocode = xxx;	**%let ocode = xxx;**

6.5.7 Finally, edit the title defined by the statement that begins '%let title='. This title will appear in all output files. The title need not include the date, because the date will automatically be included in the output files. Each SAS statement must end with a semi-colon.

2. xxxacc.sas

6.5.8 Find the where clause in the SQL select statement in this file. In the where clause, edit the line that refers to the area2 field. For example, for the GSA area, it should be:

and l.area2='GSA'

or, for a run that combines TFG and TF5475, it would read

and l.area2 in ('TFG', '5475').

Note: These lines do not end with a semi-colon or any other special character.

6.6 Run SAS Program to Identify Outliers

6.6.1 If necessary, change to the working directory. Make sure that Section 6.5 has been completed.

6.6.2 Execute the SAS program that identifies outliers by typing sas xxxrun at the UNIX prompt.

6.6.3 The SAS program in xxxrun.sas performs the following steps:

1. Retrieve data from EPDData.
2. Identify locations and analytes with enough data.
3. Identify outliers among the locations and analytes with enough data..
4. Create output files listing outliers, suitable for printing.
5. Prepare a SAS binary format file of outliers for web review.
6. Create an ascii file that tells the web review program where to find the data.

Typical runs take anywhere from 3 to 25 minutes, depending on the quantity of data processed and the load on the machine.

6.7 Review Outlier Output Files

The following files are created during Section 6.6.6:

- xxxoutl.lst
- xxxoutlf.lst
- xxxrun.lst
- xxxenough.lst
- xxxtoofew.lst
- contents.run.lst
- xxxrun.log

Note: See Appendix A for details of these output files.

6.7.1 Review xxxrun.log, the SAS log file, for error messages.

6.7.2 Review the list of outliers in the file xxxoutl.lst.

6.7.3 Inform the QC Chemist that a new set of outliers is ready for web review, specifying Site 300 or the Livermore Site. Provide the QC chemist with a printed copy of xxxoutl.lst. when requested. Inform the QC Chemist how many outliers and outlier groups there are to be reviewed (see the last entries in the columns

headed ‘grp num’ and ‘outl num’ in xxxoutl.lst). The number of outliers will be greater than the number of outlier groups if any location-analyte combination has more than one outlier.

6.8 Documentation

Write a note in the appropriate log file stating that the work has been completed.

6.9 Web Review of Outliers

The QC Chemist reviews the outliers using a SAS program accessed through a World Wide Web (Web) page. The page is among the ERD controlled-access web pages, and is available only to authorized individuals.

Web review includes the following features:

- For each outlier, a time-series plot of the outlier and its associated data is displayed.
- Along with the plot, the reviewer is presented with a dialog box offering options to accept the algorithm recommendation, or reject the recomendation for any of several possible reasons.
- The reviewer’s decisions are saved for later uploading to EPDData (Section 6.14).

The web review code permits the reviewer to revisit an outlier and change their mind any time prior to the execution of Section 6.14.

6.10 Requirements for Web Review

In order to perform web review of outliers, the QC Chemist needs the following resources:

- Authorization to access ERD controlled-access web pages
- A desktop computer with web browser software and X Windows Server software.

6.11 Performing Web Review of Outliers

6.11.1 Start the X Windows and web browser software.

6.11.2 Use the browser to access the Outlier Review web page. The current URL is http://www-gemini.llnl.gov/outliers.html. This page should always be accessible through links from the ERD contolled-access Home Page.

6.11.3 The web page offers several options. First, use the pull down menu to choose either Site 300 or the Livermore Site.

6.11.4 For convenient reference, the outlier groups are numbered. The web page has entry boxes in which to specify either a range of outlier groups (such as numbers 37 through 56) or a list of specific outlier groups (such as 5 7 19 23 45). Each outlier group is a single location-analyte combination, which may include more than one outlier. Enter the desired outlier groups, and click the ‘Run’ button.

6.11.5 Switch to the X Windows software. Several SAS windows should appear. Included is one that shows the choices that were made in the web page, and permits the reviewer to change some of the choices. Follow the instructions in this window.

There will be further activity in various windows, including some in the web browser window.

After a short time two review windows appear. These are

- A plot of data including an outlier.
- A dialog box containing information about the outlier, and a list of review decisions from which to choose.

6.11.6 For each outlier, choose a review decision. Follow the instructions in the window to record the review decision.The basic review decision is whether to accept or reject the outlier algorithm recommendation. See Section 6.12 for a description of the color coding of the outliers in the plots.

The dialog box and plot will be updated to show the next outlier. If there is more than one outlier in a single plot only the dialog box will be updated.

6.11.7 When the review is complete, inform the Outlier Custodian.

6.12 Web Review Features

The data is plotted with sample date on the horizontal axis and analyte concentration on the vertical axis. Each sample result is plotted with either a solid circle for a detection, or an open circle for a non-detection. The colors are:

Web review color codes.

Color	Meaning
Red	**New outlier**
Magenta	**Previous outlier, still an outlier**
Yellow	**Previous outlier, no longer an outlier**
Blue	**Previous outlier, was previously overridden, and is no longer an outlier.**
White	**Not an outlier**

As each outlier is presented, the reviewer is given the option to either accept or reject the software recommendation. The dialog box presents information about the outlier, a list of review options, and a default choice.

The review options are described below.

Web review codes.

Number	Code	Description
0	r	Restore to outlier algorithm status (undo a previous review decision in this run)
1	pf	Override: Poor model fit
2	nd	Override: It is an ND (Note: only appropriate for NDs that are *above* the regression line)
3	lc	Override: Low concentration, too close to the reporting limit
4	rc	Most recent sample. Do not use this option.
5	o	Override: Other reason (please enter reason in comments field)
6	so	Previous outlier, is still an outlier
7	no	Previous outlier, is no longer an outlier (the default choice)
8	sr	Previous override, still an override
9	nr	Previous override, no longer an override (the default choice)

6.12.1 To accept the default recommendation, press the F3 key.

6.12.2 To reject (override) the default recommendation, type an override code (listed below) in the appropriate entry field. The code may be either numeric or character. There is an additional entry field for optional comments. Press F3 when ready to move to the next outlier.

- Option 0, “Restore to outlier algorithm status” should be used only when revisiting an outlier and deciding that the first review decision (in this outlier run) was incorrect. In this case, use this option to return to the algorithm recommendation. Never use this option when first reviewing an outlier.
- Options 1, 2, and 3 present the reviewer with three frequently used reasons for overriding the algorithm recommendation.
- Option 4 applies to outliers that are the most recent sample within their location-analyte group. Such outliers are presented for information only. Do not use this option: where applicable it will be automatically selected by the software. However, these points are presented in order to bring them to the reviewer’s attention. Do not override such outliers. They will be presented again in a future outlier run, if they are still outliers after additional data has been acquired.
- Option 5 offers the reviewer the opportunity to override an algorithm recommendation for some other reason, and enter a brief reason.
- Option 6, “Previous outlier, still an outlier” is used when a previous outlier run declared the point to be an outlier and the current run erroneously declares

that it is not an outlier. Option 6 is used to override points for which the algorithm recommends option 7.

- Option 7, "Previous outlier, no longer an outlier" is recommended by the algorithm when a point that was found to be an outlier in a previous run is no longer found to be an outlier. This option is the default; it is not necessary to enter '7' or 'no' in the dialog box. Just press F3.

- Option 8, "Previous override, still an override" is used when a previous outlier run declared the point to be an outlier, the reviewer overrode the algorithm and declared that it is not an outlier, the current run finds that it is no longer an outlier (thus not needing to be overridden anymore), but the reviewer decides that it should still be considered an overridden outlier. This option is used to override points for which the algorithm recommends option 9. This option is rarely used.

- Option 9, "Previous override, no longer an override" is recommended by the algorithm when a previous outlier run declared the point to be an outlier, the reviewer overrode the algorithm and declared that it is not an outlier, and the current run finds that it is no longer an outlier (thus not needing to be overridden anymore). This option is the default; it is not necessary to enter '9' or 'nr' in the dialog box. Just press F3.

6.13 Upload, Update, and Append Outliers

After review is complete, the information is incorporated into EPDData. The Outlier Custodian performs the following steps:

1. Upload to an EPDData working table, either uo3_statrecs (Site 300) or uol_statrecs (Livermore Site). See Section 6.14 and Section 6.15.

2. Update the stat_code, stat_flag, and stat_date fields in the global analysis table. See Section 6.16.

3. Append the working table to the global statrecs table. See Section 6.16.

4. Delete all rows in the working table. See Section 6.16.

Note: Until the update and append steps (Section 6.16) are initiated, nothing is final. It is still possible, for example, for the reviewer to revisit some outliers and change his or her mind about them. If this takes place, then the upload step (Section 6.14) will have to be repeated. In fact, any time before Section 6.16 is performed it is possible to return to any step in Sections 6.4 through 6.12. Of course, all subsequent steps would have to be repeated.

6.14 Upload to EPDData

6.14.1 Change to the working directory (see Section 6.4).

This procedure uses two working tables, uol_statrecs (LivermoreSite) and uo3_statrecs (Site 300). Access to these tables is limited to individuals authorized to peform this procedure.

6.14.2 Manually verify that the working is empty. If it is not empty, then there has been a deviation from normal procedure. If this is the case, do not proceed. See Section 6.16.3 for further instructions.

6.14.3 Run the SAS program xxxupl.sas using the UNIX command sas xxxupl. This program uploads the current outliers to the appropriate EPDData working table. (The working table is specified using the wtable macro variable in the autoexec.sas file; see Section 6.5)

6.14.4 As the program runs it will print a summary table of the review results. Use this summary to make sure that all outliers were reviewed. Note the number of outliers uploaded to EPDData, for comparison with Section 6.16. The number of outliers uploaded will usually be fewer than the number identified, because outliers that are the most recent sample are not uploaded. Review the output files described in Section 6.15.

6.14.5 A single warning message "WARNING: 0 rows processed." in the SAS log file is normal; it indicates that the working table was empty prior to uploading, as it normally should be (see Section 6.16). This warning message will not appear if the working table contained data. See Section 6.16.3 for a discussion of how this might occur.

6.15 Review Upload Output Files

The following output files are created by the upload process:

- xxxrev1.lst
- xxxrev2.lst
- xxxrev3.lst
- xxxrev4.lst
- xxxwrk1.lst
- xxxwrk2.lst
- chkuo3.before and chkuo3.after (Site 300), or chkuol.before chkuol.after (Livermore Site)
- contents.upl.lst

Note: See Appendix A for details of these output files.

6.15.1 Make sure that the reviewer reviewed all outliers. This can be checked either of two ways: (1) from the output that is displayed as the program xxxupl.sas runs (instructions are given), or (2) by inspecting file xxxrev4.lst. The first table in.lst gives the frequency of each review decision. A blank indicates no review, so if the column labeled 'REASON' has any blanks in it then the reviewer has not reviewed all outliers. Tell the reviewer to finish Section 6.11.

6.16 Update and Append

6.16.1 Log in to EPDBS and change to the working directory (see Section 6.4).

6.16.2 Run the Monitor routine S1 T2. This routine should take at most a few seconds to run.

The T2 routine does three things:

1. It appends the records in the working table to the global statrecs table.
2. It updates matching records in the analysis table.
3. If steps 1 and 2 were successful it empties the working table.

6.16.3 Examine the log file created by the T2 routine, and check that the number of records updated and appended matches the number uploaded, as indicated in files xxxwrk2.lst and chkuol.after (or chkuo3.after, as appropriate). The log file is named stat3_rowcnts.log (Site 300), or statl_rowcnts.log (Livermore Site).

If the working outliers table is not empty it indicates one of two possible situations:

1. The previous outlier run is not complete. In particular, Section 6.16 was not performed. First check the appropriate log file (S3-outlier-log or LS-outlier-log). If the log file indicates that the step was performed, then check the values of loc_id to see what AREA2 they belong to. Do not perform Section 6.14 until either Section 6.16 for the previous run is performed, or it is determined that it is not necessary to update and append the records in the working table.
2. Section 6.14 of this outlier run was previously performed. Check the appropriate log file. This may have occurred if errors in the run were found, or if the reviewer changed his or her mind, after Section 6.14 was performed the first time. Since this is a previous upload from this run, it is acceptable to replace it. Proceed with Section 6.14; the records will be deleted from the working outliers table before the upload takes place.

7.0 QA RECORDS

A printed copy of the following files should be retained as QA records:

7.1 xxxrun.sas

7.2 xxxacc.sas

7.3 xxxupl.sas

7.4 autoexec.sas

7.5 xxxrun.log

7.6 xxxupl.log

7.7 xxxrun.lst

7.8 xxxoutl.lst

7.9 xxxwrk1.lst

7.10 xxxwrk2.lst

7.11 xxxrev1.lst

7.12 xxxrev2.lst

7.13 xxxrev3.lst

7.14 xxxrev4.lst

7.15 chkuo3.sql, chkuo3.before, chkuo3.after (Site 300)

7.16 chkuol.sql, chkuol.before, chkuol.after (Livermore Site)

8.0 ATTACHMENT

Attachment A—Output Files

Attachment A

Output Files

Attachment A

Output Files Created During Outlier Run

A-1 Output Files from Section 6.3.3:

The following output files are created as outliers are identified during the outlier identification step:

- xxxoutl.lst
- xxxoutlf.lst
- xxxrun.lst
- xxxenough.lst
- xxxtoofew.lst
- contents.run.lst

A-1.1 Output File xxxoutl.lst

This is the primary output file for Section 6.3.3. It contains information about each outlier that was identified:

grp num. A unique number for each location-analyte combination with an outlier.

outl num. A unique number for each outlier identified in this run.

prev code. The outlier code (if any) from the previous outlier run.

new code. The (new) outlier code determined by this outlier run.

loc id. The location (loc_id in the EPDData sample table).

chem id. The analyte (description in the EPDData parameter table).

sample date. The sample date (sampled in the EPDData sample table).

los ind. Reporting limit indicator (‘<‘ if the result is less than the reporting limit).

result. Measured concentration (result in the EPDData analysis table).

units. Units of measure for result.

lab. The analytical laboratory that performed the analysis.

req anal. The requested analysis code.

qa flag. The clp_qa_flag from the EPDData analysis table.

A-1.2 Output File xxxoutlf.lst

This file contains a table that lists how many outliers have changed status:

prevcode. The outlier code (if any) from the previous outlier run.

newcode. The (new) outlier code determined by this outlier run.

change. Should always contain 'y' for yes.

count. The number of outliers that have changed from prevcode to newcode. For example, if in the first row of this table prevcode is blank, newcode is 'oh', and count is 192, then 192 new high outliers were found ('oh' for high outlier, 'ol' for low outlier, 'xx' for override).

A-1.3 Output File xxxrun.lst

This file contains information about the database from which data was retrieved (normally a gemini database) and may contain other miscellaneous information:

obs. Not useful (row number of the printed output).

gemnum. The gemini number (1 for gemini1, 2 for gemini2**)**.

gemdate. The gemini copyover date.

gemtitle. A title that is used in the output files.

host. The computer on which the database resides.

db. The name of the database.

A-1.4 Output File xxxenough.lst

This file lists information about location-analyte combinations for which there was enough data to look for outliers:

loc_id. Sample location.

chem_id. Analyte name.

nnd. Number of non-detects.

nhit. Number of detections.

ntot. Total number of samples.

beg_date. Earliest sample date (at this location for this analyte).

end_date. Most recent sample date (for which analytical results are in the database).

min_los. Minimum reporting limit among the analyses (at this location for this analyte).

max_los. Maximum reporting limit among the analyses.

min_res. Minimum result among the analyses.

max_res. Maximum result among the analyses.

A-1.5 Output File xxxtoofew.lst

This file lists information about location-analyte combinations for which there was not enough data to look for outliers:

loc_id. Sample location.

chem_id. Analyte name.

nnd. Number of non-detects.

nhit. Number of detections.

ntot. Total number of samples.

beg_date. Earliest sample date (at this location for this analyte).

end_date. Most recent sample date (for which analytical results are in the database).

min_los. Minimum reporting limit among the analyses (at this location for this analyte).

max_los. Maximum reporting limit among the analyses.

min_res. Minimum result among the analyses.

max_res. Maximum result among the analyses.

For example, the run labeled 'b83.98a' found a number of outliers from the late 1980's. One might expect these to have been identified during the first 'b83' run (because the first 'b83' run took place in 1997). By comparing b83enough.lst from the 1998 b83 run with b83toofew.lst from the 1997 b83 run, one finds that the new outliers are in datasets that in 1997 did not yet have enough data for an outlier run.

A-1.6 Output File contents.run.lst

This file lists the variable names in some of the SAS datasets created during an outlier run. These technical details are for reference only, and should not need to be reviewed during normal outlier runs.

A-2 Output Files from SOP 5.21, Sections 6.5.1 and 6.5.3

The following output files are created during the Upload and Append steps (Sections 6.5.1 and 6.5.3):

- **xxxrev1.lst**
- **xxxrev2.lst**
- **xxxrev3.lst**
- **xxxrev4.lst**
- **xxxwrk1.lst**
- **xxxwrk2.lst**
- **contents.upl.lst**
- **chkuol.before (Livermore Site) or chkuo3.before (Site 300)**
- **chkuol.after (Livermore Site) or chkuo3.after (Site 300)**

A-2.1 Output File xxxrev1.lst

This file contains information about the outlier review process. It is the primary review file. Only outliers that have been reviewed are included. Compare the number of records listed here with that in xxxoutl.lst.

grp num. A unique number for each location-analyte combination with an outlier in this run. The same as in xxxoutl.lst.

outl num. A unique number for each outlier identified in this run. The same as in xxxoutl.lst.

loc id. The location (loc_id in EPDData).

chem id. The analyte (description in the EPDData parameter table).

samp date. The sample date (sampled in the EPDData sample table.

los ind. Reporting limit indicator (< if the result is less than the reporting limit).

result. Estimated concentration (result in the EPDData analysis table).

units. Units of measure for result.

prev code. The outlier code (if any) determined by the previous outlier run.

new code. The (new) outlier code determined by this outlier run.

rev code. The review decision for each outlier (for example, 'xx' to override the algorithm).

stat code. The outlier code that will be uploaded to Ingres, where it will become the stat_code field in the Ingres working table.

up ld. A flag indicating whether ('y') or not ('n') this record will be uploaded to Ingres.

lab. The analytical laboratory that performed the analysis.

req anal. The requested analysis code.

qa flag. The clp_qa_flag from the EPDData analysis table.

Note: The rev code determines whether the stat code is equal to the new code, the prev code, or the override code (the override code is always 'xx').

A-2.2 Output File xxxrev2.lst

This file contains the same records in the same order as xxxrev1.lst, but with additional detail concerning the review decisions. Only outliers that have been reviewed are included.

grp num. As in xxxrev1.lst

outl num. As in xxxrev1.lst

prev code. As in xxxrev1.lst

new code. As in xxxrev1.lst

rsn. The reason code entered by the review during the web review process.

rev code. As in xxxrev1.lst

stat code. As in xxxrev1.lst

up ld. As in xxxrev1.lst

prev flag. The value of the stat_flag field of the EPDData analysis table previous to this outlier run.

new status. A description of the new status of this outlier.

review status. A description of the review decision.

A-2.3 Output File xxxrev3.lst

This file has the same columns as xxxrev2.lst, but contains only those outliers that will not be uploaded to Ingres. At present, the only reason for not uploading is the status "Most recent."

A-2.4 Output File xxxrev4.lst

This file contains a frequency table of review decisions:

reason. The reason code entered by the reviewer during the web review process. *Important: if any row has a blank reason, then the web review process was not completed.*

Frequency. The number of outliers that were given each reason code.

Cumulative frequency. The sum of the frequencies.

A-2.5 Output File xxxwrk1.lst

This file contains a complete list of the data (all fields and all records) that will be uploaded to the Ingres working table. Where the SAS variable name is different than the Ingres field name, both names are given in the column header.

A-2.6 Output File xxxwrk2.lst

This file contains two tables with counts of outlier types. The first table is generated from the SAS dataset prior to uploading. The second table is derived directly from the Ingres working table. The tables should agree exactly.

A-2.7 Output File chkuol.before (Livermore Site) or chkuo3.before (Site 300)

These files contain output from the SQL query in chkuol.sql (or chkuo3.sql). This query is run before uploading to Ingres, and can be used to find out what records, if any, were in the working table before uploading. Normally there should be none, but if for some reason the upload is repeated, there may be some. See Section 0.

A-2.8 Output File chkuol.after (Livermore Site) or chkuo3.after (Site 300)

These files contain output from the SQL query in chkuol.sql (or chkuo3.sql). This query is run after uploading to Ingres.

Both queries (before and after) are run independently of SAS, so that the 'after' output can be used as an independent check on whether SAS correctly uploaded the data.

A-2.9 Output File contents.upl.lst

This file lists the variable names in some of the SAS datasets created during an outlier run. These technical details are for reference only, and should not need to be reviewed during normal outlier runs.

A-3 Outlier Data Dictionary

This section describes the contents of the fields in the EPDData statrecs table, and the outlier related fields in the EPDData analysis table. The fields in the uol_statrecs and uo3_statrecs tables are the same as the statrecs table.

The first time that an outlier is identified it is appended to the statrecs table. Over time, additional data may indicate that an outlier is no longer an outlier. In this case, another record for that outlier is appended to the statrecs table (with a later stat_date), and the outlier fields in the analysis table are updated to reflect the new status. The end_date and result_4 fields in the statrecs table should reflect the fact that additional data was acquired (this is why they are included in the table).

A-3.1 Outlier Review Codes:

stat_code. The outlier review codes are stored in the statrecs and analysis tables.

ol. low outlier.

oh. high outlier.

xx. Identified by the algorithm as a possible outlier, but the reviewer decided it is not.

A-3.2 Analysis Table Data Dictionary

The EPDData analysis table includes three outlier related fields:

- **stat_flag.** A count of the number of times the outlier appears in the statrecs table. Equivalent to the number of times it has changed outlier status.
- **stat_date.** The most recent stat_date in the statrecs table for this outlier.
- **stat_code.** Same as stat_code in the statrecs table.

A-3.3 Statrecs Table Data Dictionary

This table contains a record of all outliers found by the algorithm and the review decision that was made for each. It includes results from all outlier runs, including both Site 300 and the Livermore Site.

- **loc_id.** The sample location (loc_id from the EPDData sample table).
- **sampled.** The sample date (sampled from the EPDData sample table).
- **parameter.** The analyte parameter code (chem_id from the EPDData analysis table).
- **req_analys.** Requested analysis (req_analys from the EPDData sample table).
- **log_no.** Sample log number (log_no from the EPDData sample table).
- **result.** Analytical result (result from the EPDData analysis table).
- **stat_code.** A code representing the type of outlier. See below for a definition of the codes.
- **result_1.** The outlier group number (grpnum).
- **result_2.** The studentized residual value for the outlier. This is the number of standard deviations the outlier is away from the regression line.
- **result_3.** The standard deviation used in calculating result_2.
- **result_4.** The number of samples in the set of data that includes the outlier.
- **remarks.** A brief description of the outlier status. In most cases, this is used to give a reason why the reviewer overrode the outlier algorithm.
- **project.** The project field from the EPDData sample table.
- **updater.** The updater field from the EPDData sample table.
- **stat_date.** The date and time when the outlier run was performed.
- **begin_date.** The earliest sample date in the set of data that includes the outlier.
- **end_date.** The most recent sample date in the set of data that includes the outlier.
- **analyzed.** The date the sample was analyzed (analyzed from the EPDData analysis table).
- **gemdate.** The copyover date of the gemini database that provided the data.

A-4 Outlier Identification Logic

The outlier identification program uses a simple statistical method to identify possible outliers. For each location-analyte combination, a simple linear regression model is fit to the data. Non-detections are included, using the value of the reporting limit. Any analytical more than three standard deviations away from the regression line is considered a possible outlier.

This approach is simple and easily understood. However, it is not perfect. For example, analyte trends over time are not necessarily linear. Also, unusually large detection limits can unduly influence the regression line. Therefore, human review of the possible outliers is an essential part of the algorithm.

Outlier status is intended to warn data users that a result is inconsistent with the remainder of the data (location-analyte combination) with which it is associated. Users should never blindly reject data solely on the basis of outlier status.